SESSION

"Best Practices for High Volume Web Sites"

NUMBER

Presentation Number PE420511, E420521

SPEAKER

Dr. Willy Chiu

TITLE

VP, High Volume Web Sites

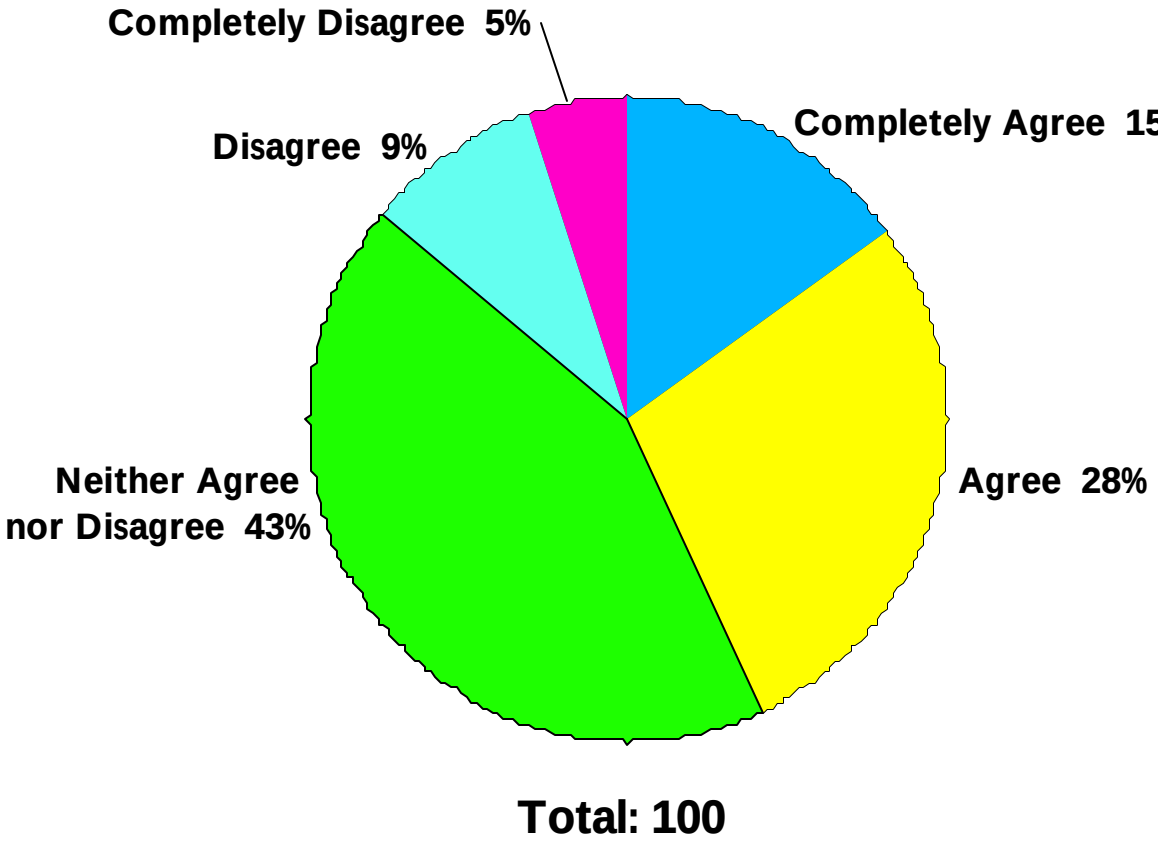
AFFILIATION

IBM

wchiu@us.ibm.com

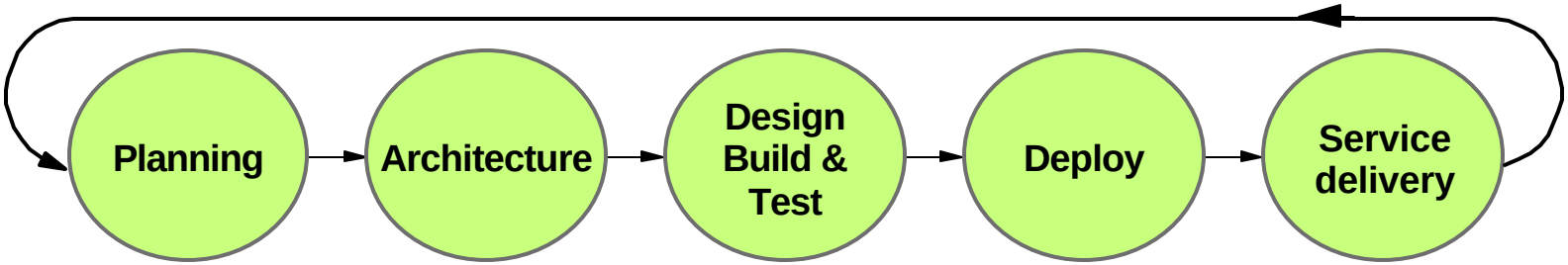
I/T Professional's Survey

(Marketplace Needs to Improve the Performance of the Web Site?)



Source: Zona B2B Marketplace Study - 4Q2000 (Copyright 2001, Zona Research)

High Volume Web Sites
Best Practices



Life Cycle of a Web Site

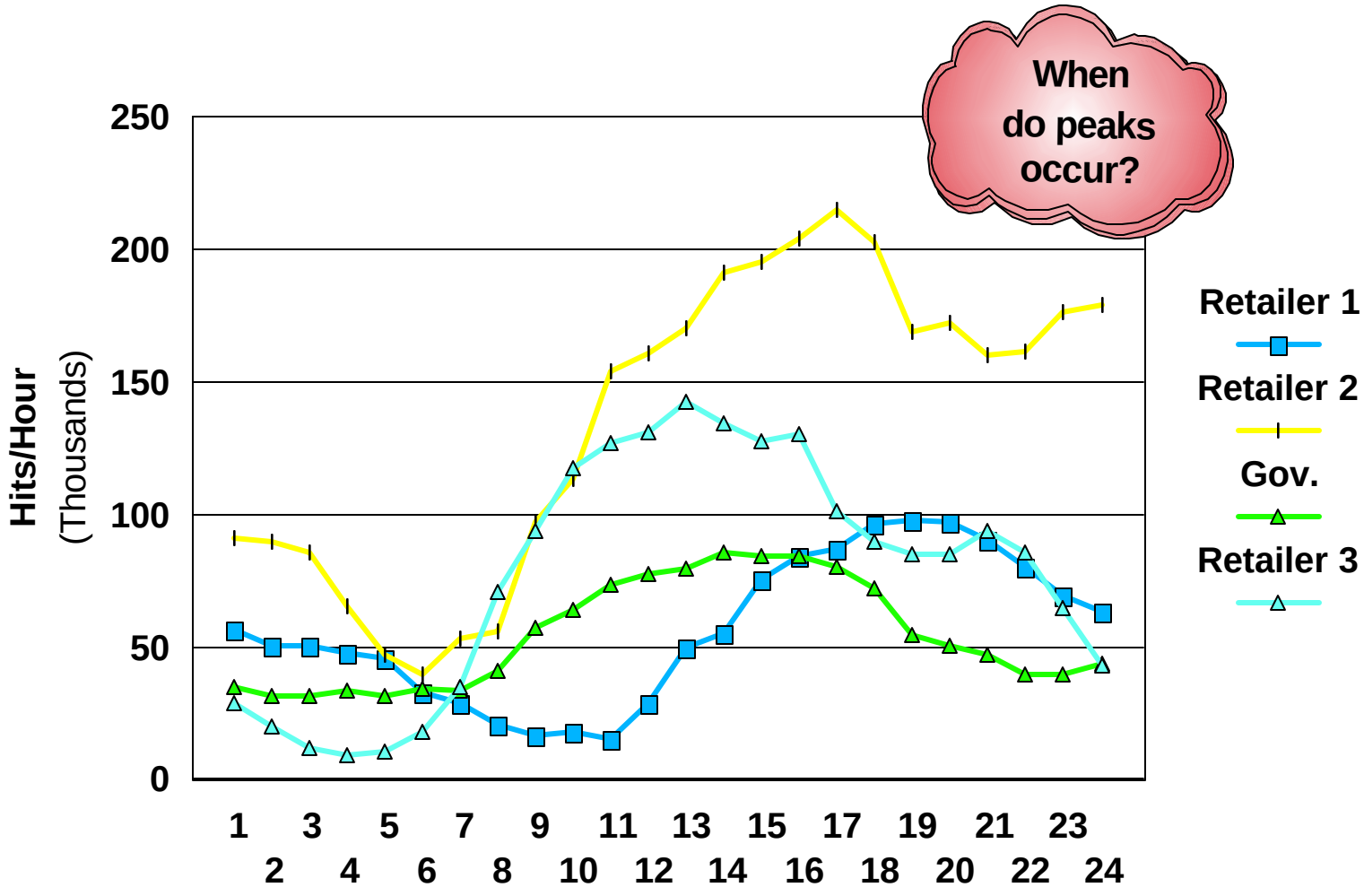
Best Practices

- **Design for scalability**
- **Manage end-to-end performance**
- **Plan for growth**
- **Design pages for performance**
- **Know your workload**

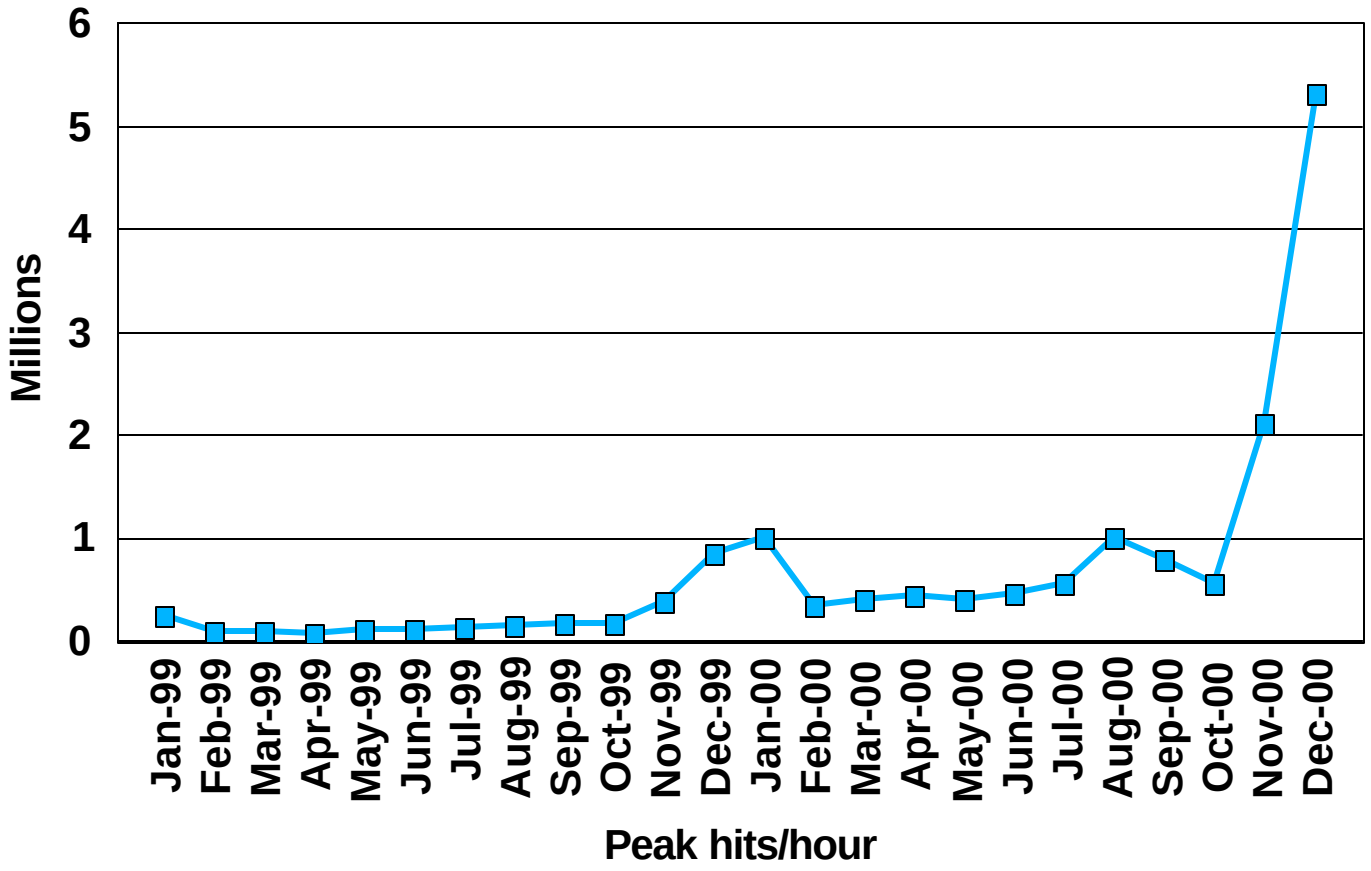
Note: Material regarding the topics addressed in this presentation may be obtained at:
"www.ibm.com/developerworks/library/hvws"

Patterns for e-Business & Workload Taxonomy

Work Site Loads
Usage Pattern Over One Day

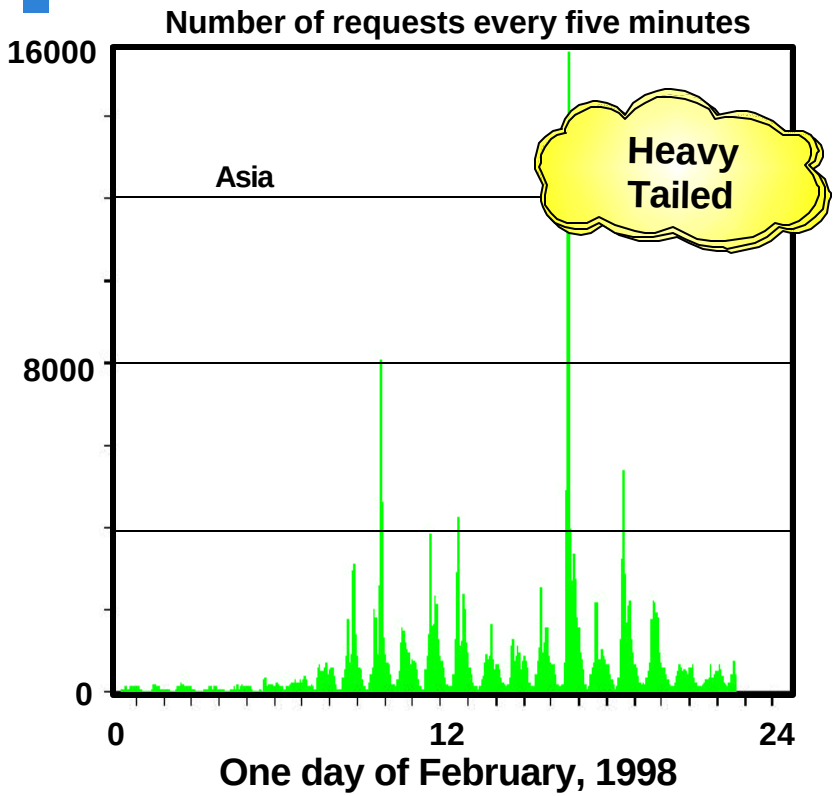


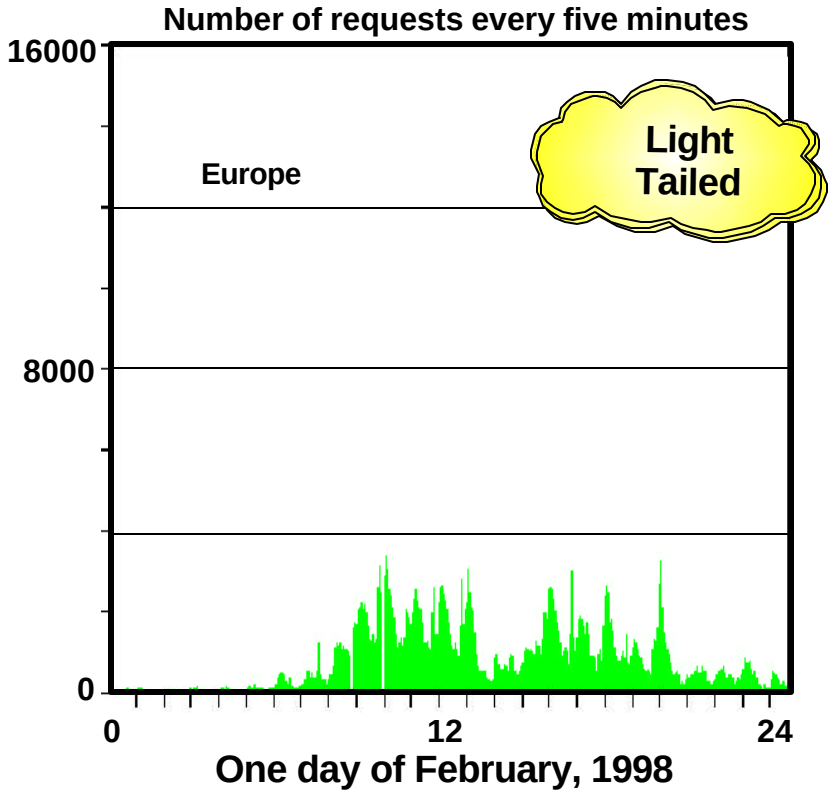
Online Retailer
Monthly Trend in Peak Hits/Hour
Seasonal Effects



Bursty Traffic Patterns

1998 Nagano Olympic Games





Contains very large bursts

Large traffic bursts - strong interest in Japan for events related to ski jumping

Much less bursty

Traffic more scattered due to time differences and broader interests in events

Sydney Olympics much less bursty due to 18 hour broadcast delay

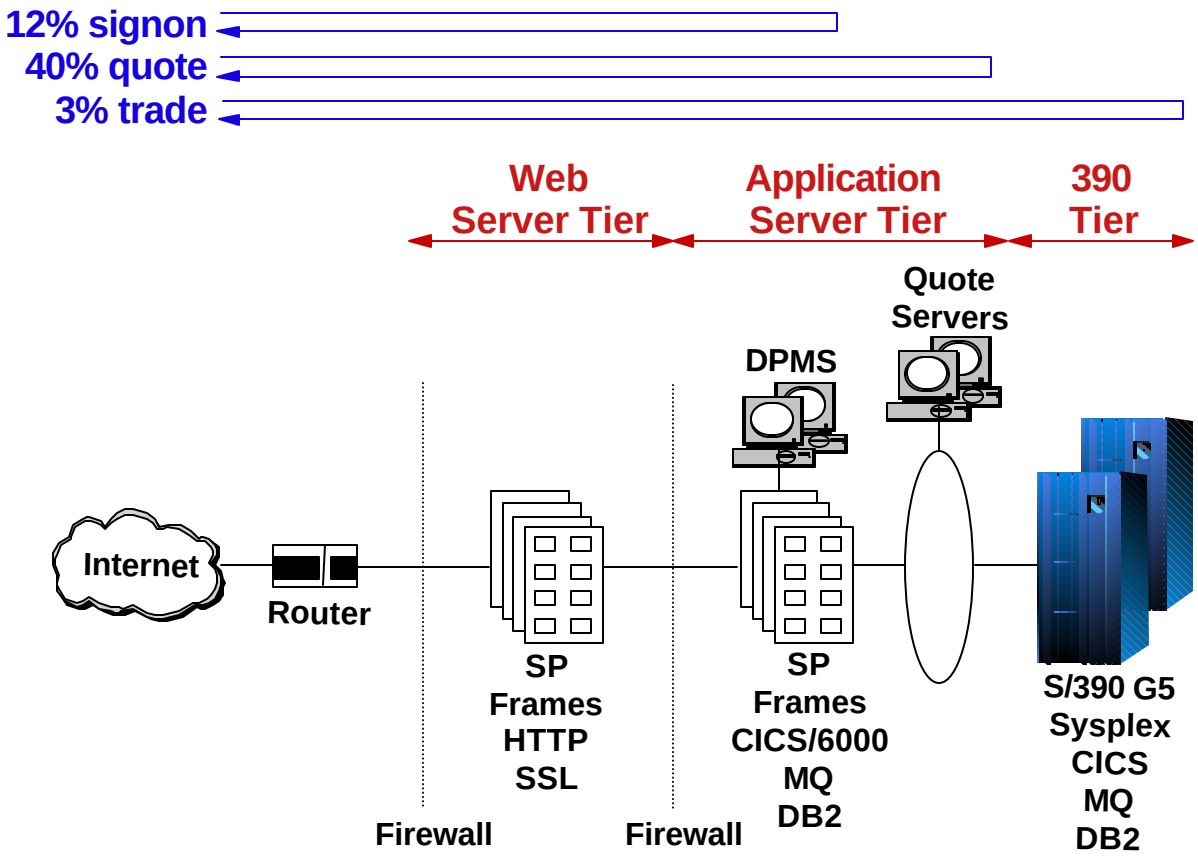
Design for Scalability
Best Practices

Six steps to scaling:

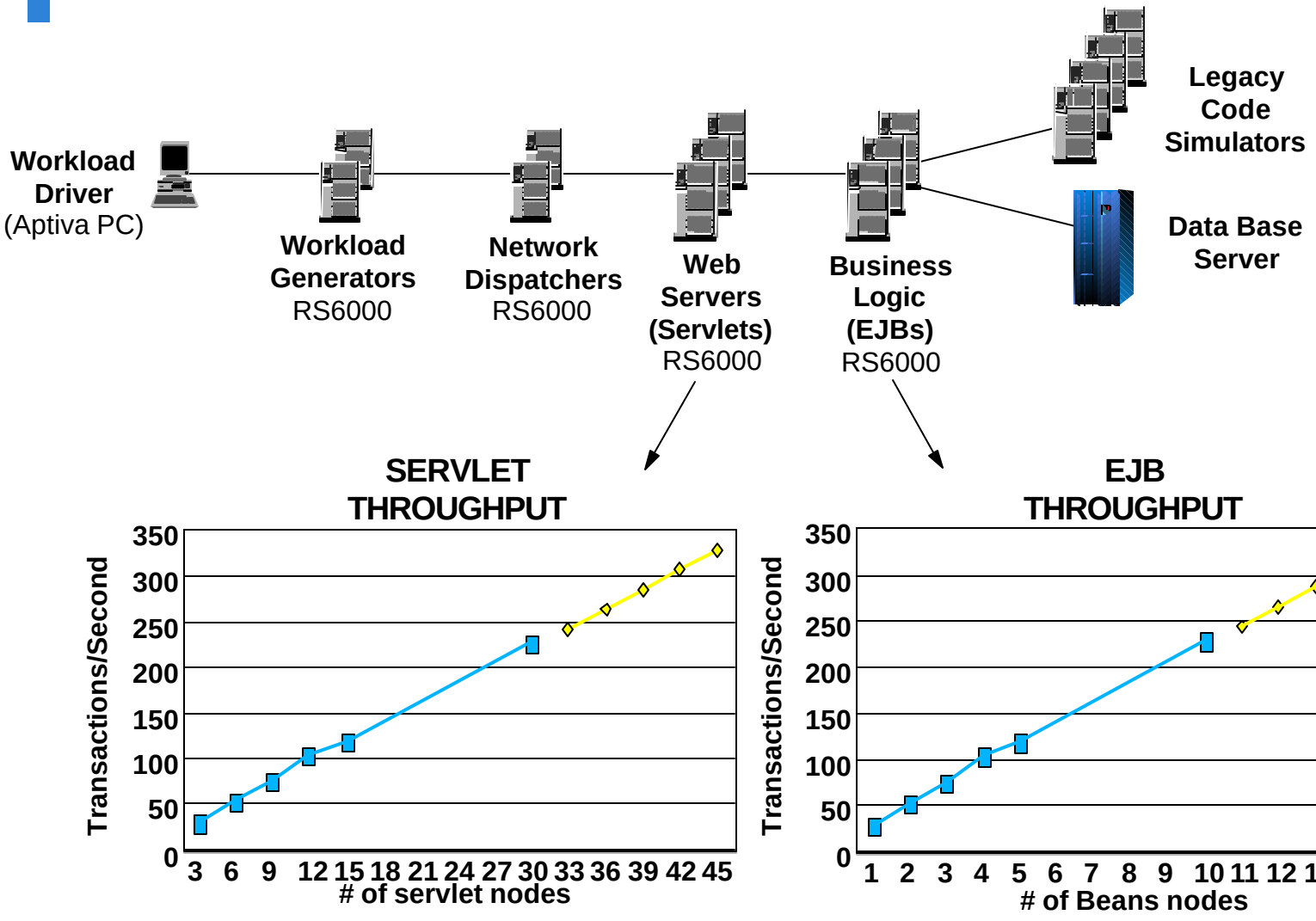
- Understand the environment
- Categorize your workload
- Determine the components most impacted
- Select the scaling techniques to apply
- Apply the techniques
- Reevaluate

Highly Scalable
Online Stock Trading Web Site

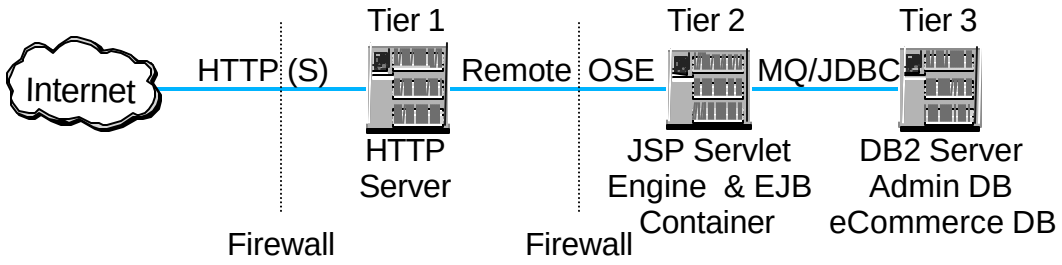
1/00	Secure site
110K	Peak concurrent users
500K	Peak trades/day
10.7M	Peak transactions/day
80M 4000	Peak hits/day Peak trans/sec

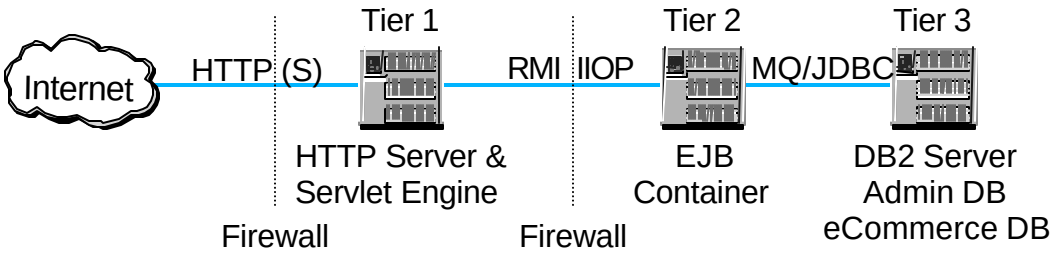


Large Scale WebSphere EJB/Servlet Configuration



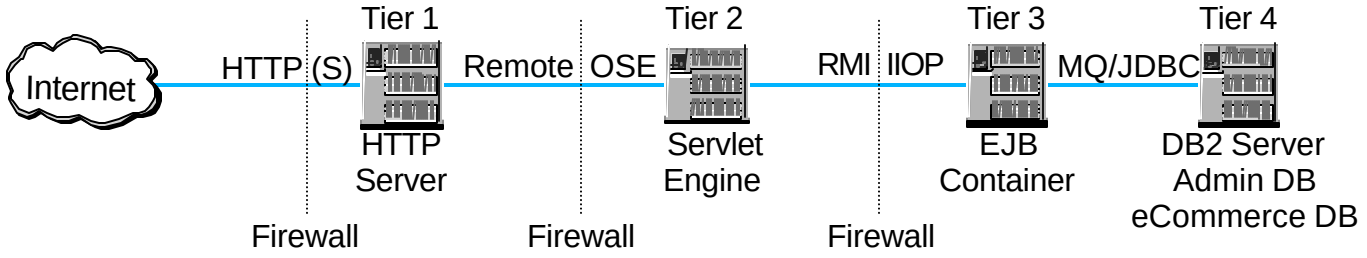
WebSphere Architecture Alternatives



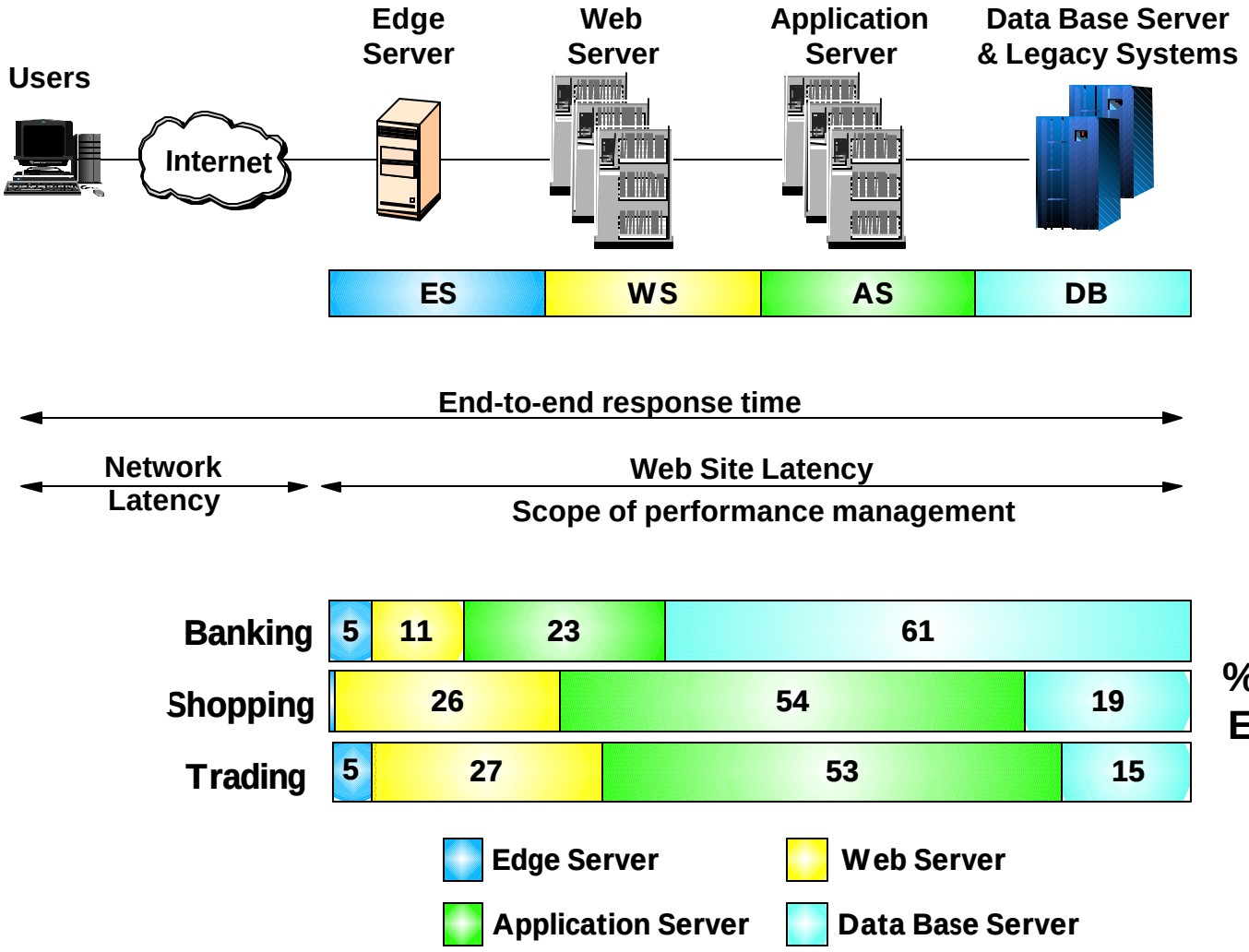


Architectural choices typically driven by:

- **Security, isolation from the open Internet**
- **Geographical considerations - separate data centers**
- **Separation of function (presentation and business logic)**



Where do you focus your scalability and tuning ?



Workload Characteristics for Scaling

System Workload Characteristics	Publish/ Subscribe	Online Shopping	Customer Self-Service	Trading	Business-To-Business
Transaction Volumes	○	◑	◑	●	◑
Dynamic Content	◑	◑	◑	●	◑
Dynamic Searches	○	●	○	○	◑
User Specific Responses (Personalization)	○	○	◑	●	◑
Cross-session Information	○	●	●	●	●
Transaction Complexity	○	◑	●	●	●
Legacy Integration	○	◑	●	●	●
Transaction Volume Swings	○	◑	◑	●	●
Data Volatility	○	○	○	●	◑
Number of Content Publishers/Sources	●	○	○	●	◑
Page Content Volatility	●	●	◑	●	◑
Number of Unique Items per page	●	◑	○	◑	◑
Number of Page Views	●	◑	○	◑	◑
Percentage of Secure Pages	○	◑	◑	●	●
Security, Authentication etc.	○	●	●	●	●

○ Low ◑ Medium ● High

Refer to "Design for Scalability" White Paper for further information
"www.ibm.com/developerworks/library/hvws"

How Do Workloads Dynamics Impact Solution Components?

System Workload Characteristics	Web Servers	Web Appl Servers	Network	Security Servers	Fire-walls	Ent. Inf. Systems	Data Base Server
Transaction volumes	○	○	○	◑	○	●	●
Dynamic content	●	◑	◑	◑	○	○	○
Dynamic searches	●	●	◑	○	◑	◑	●
User specific responses (personalization)	○	●	○	○	○	●	●
Cross-session information	◑	●	○	○	○	○	◑
Transaction complexity	○	●	○	◑	○	●	●
Legacy integration	○	●	◑	◑	○	●	●
Transaction volume swings	◑	●	○	○	○	●	●
Data volatility	●	●	○	○	○	◑	●
Number of content publishers/sources	●	○	●	○	○	○	○
Page content volatility	●	●	◑	○	○	○	●
Number of unique items per page	○	◑	○	○	○	●	●
Number of page views	●	○	●	○	●	○	○
Percentage of secure pages	●	○	○	●	○	○	○
Security, authentication, etc.	●	●	○	●	○	●	○

○ Low ◑ Medium ● High

Refer to "Design for Scalability" White Paper for further information
"www.ibm.com/developerworks/library/hvws"

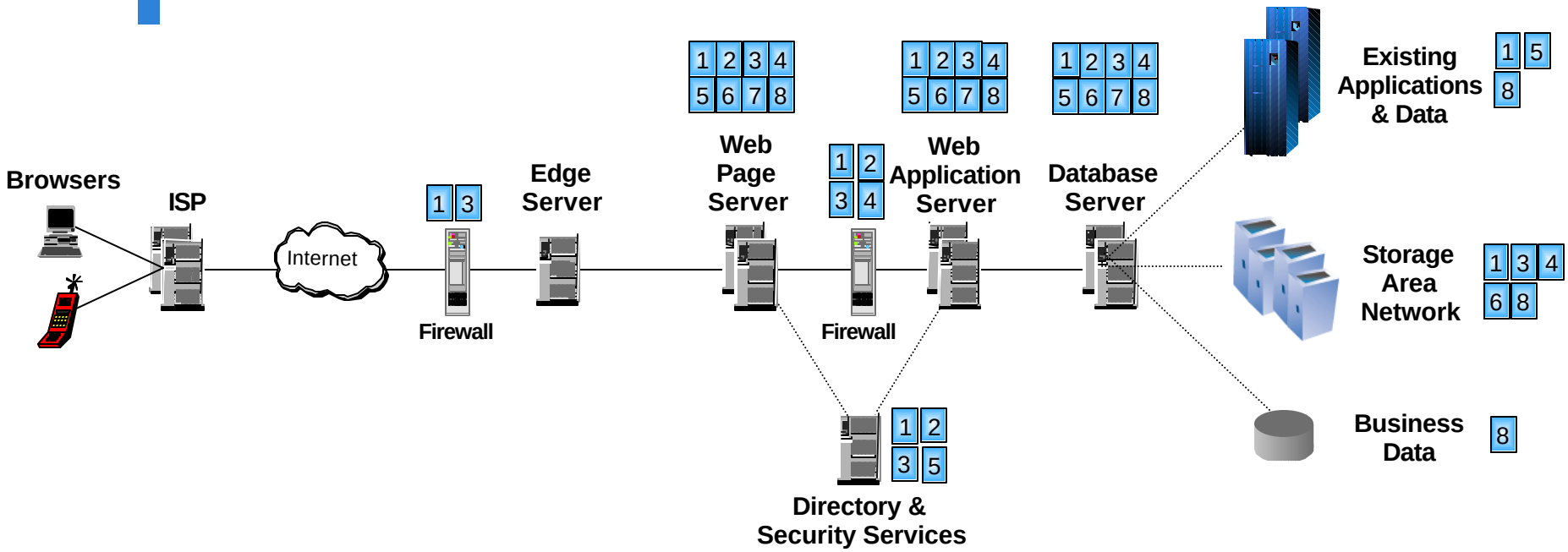
How Do Scaling Techniques Relate to Scaling Objectives ?

Scaling Technique	Increase Power	Improve Efficiency	Shift / Reduce Load
Faster Machine	x		
Replicate Machines	x		
Specialized Machines	x	x	
Segmented Workload		x	x
Request Batching		x	
User Data Aggregation		x	
Connection Management		x	
Caching		x	x

All of these techniques are in use somewhere, yet our top customers are not using all of the techniques that would be useful.

Refer to "Design for Scalability" White Paper for further information
"www.ibm.com/developerworks/library/hvws"

Design for Scalability Best Practices



Techniques

1. Faster Machine
2. Replicate Machines
3. Specialized Machines
4. Segmented Workload
5. Request Batching
6. User Data Aggregation
7. Connection Management
8. Caching

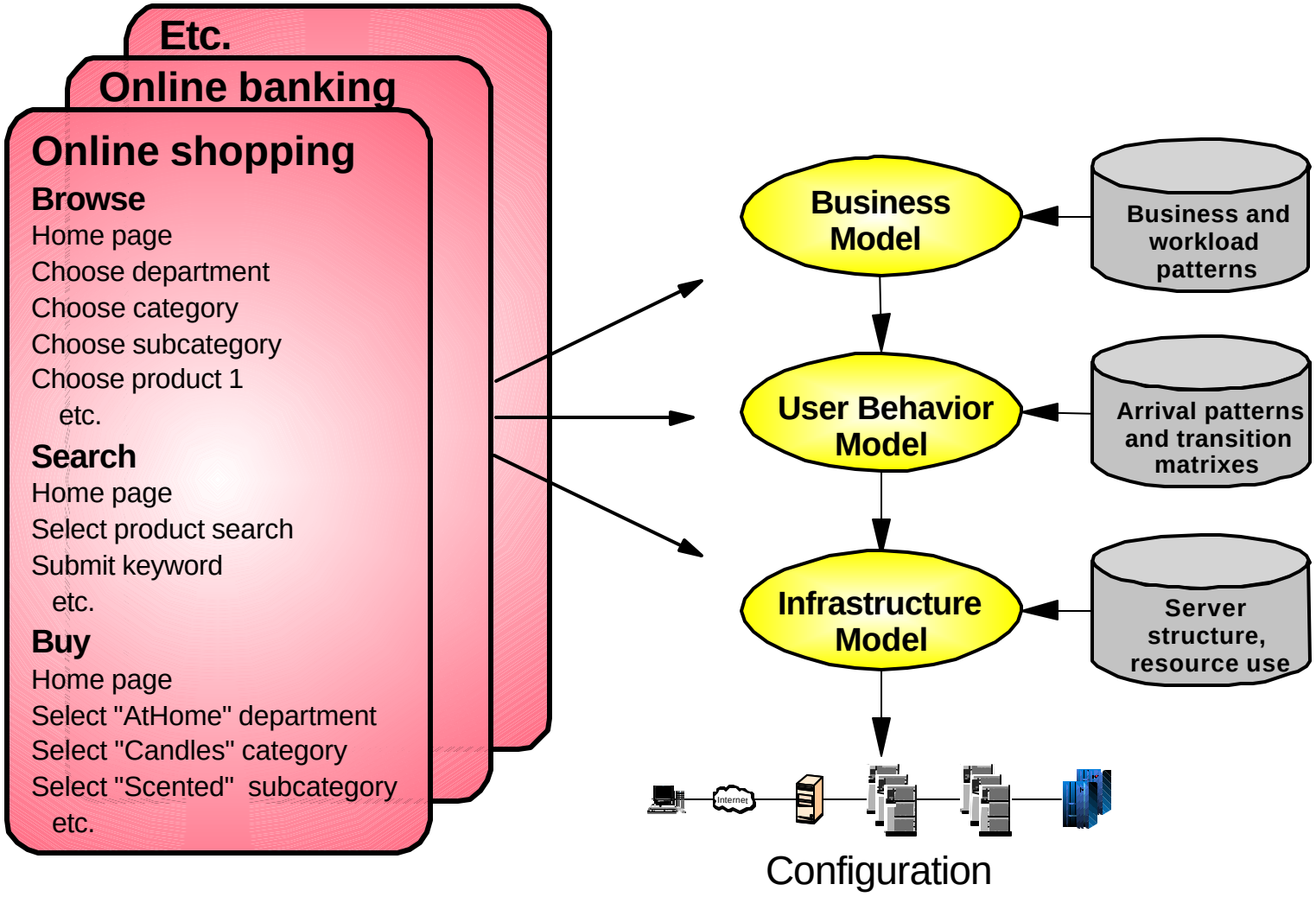
Best practice: Follow a methodology to optimize for scalability

- Understand thoroughly your application environment
- Categorize your workload
- Determine the components most impacted
- Select the scaling techniques to apply
- Apply the techniques

Planning for Growth
Lessons Learned

- Determine your workload characteristics
  - Trends
  - Burstiness
- Plan your infrastructure from end to end
  - Leads to balanced solutions
- Integrate capacity planning into your I/T and business processes

Capacity Planning Process with the HVWS Simulator



Web Page Measurement Criteria

Average server response time	**<0.5 sec**
Number of items per page	**<20**
Page load time	**<30 sec ***
Page size in bytes	**<64K**

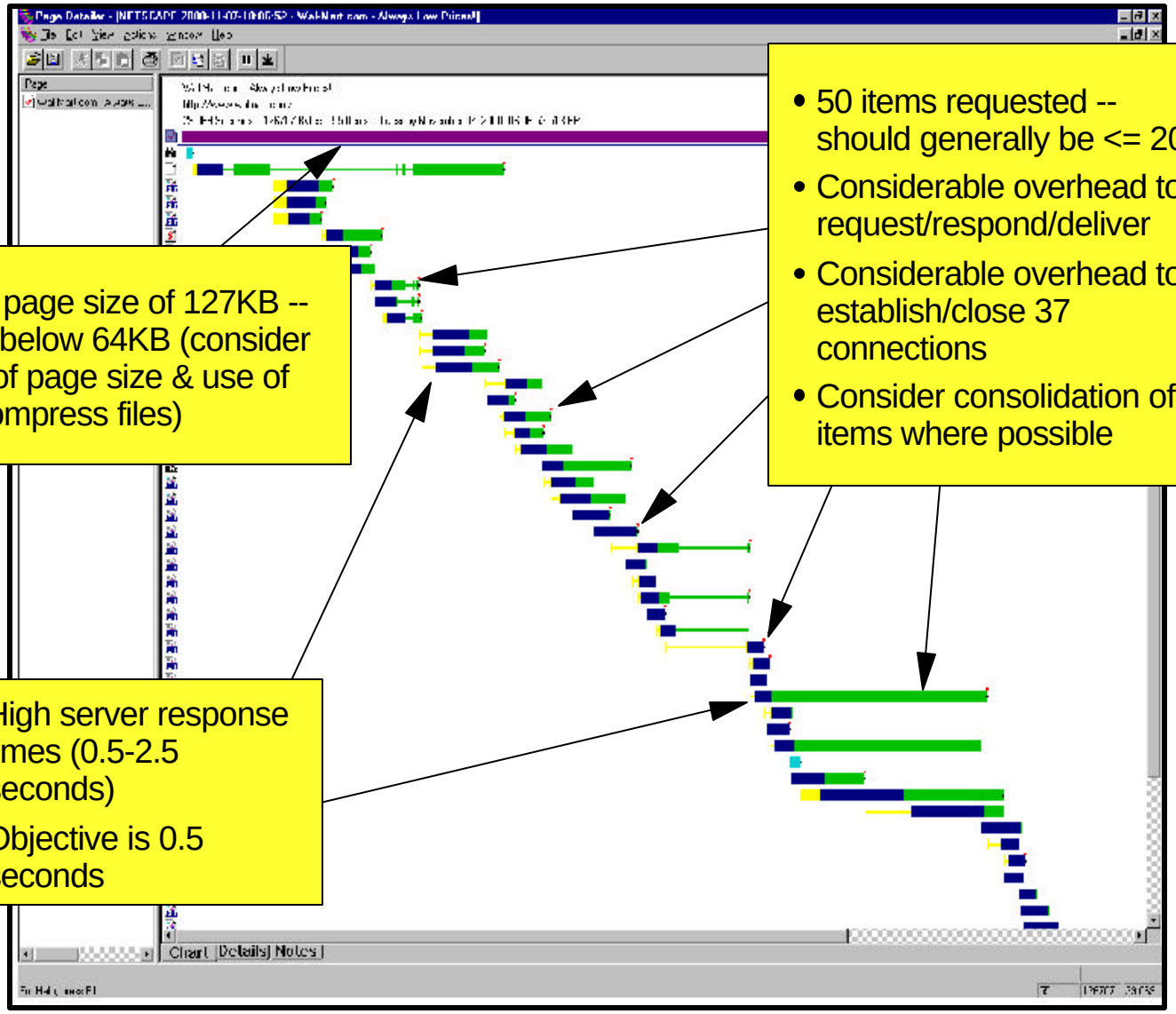
Seconds to load *	Ranking
Less than 10	Excellent
10-15	Very good
15-20	Good
20-25	Adequate
25-30	Slow
Over 30	Unacceptable

* Based on 56K connection speed

- **Minimize number, size, and complexity of page items -- consider the business value of each**
- **Store and retrieve items used more than once from the browser cache**
- **Avoid server redirection, use persistent connections, and use same code level across servers**
- **Avoid unnecessary use of SSL, banner ads, and dynamic pages**
- **Use preproduction utilities that remove extra white space from the source HTML (e.g., GZIP)**
- **Structure sequences to fully exploit parallelism - request items early**

Refer to "Design Pages for Performance" White Paper for further information
"www.ibm.com/developerworks/library/hvws"

WebSphere Studio PageDetailer Analysis



- Excessive page size of 127KB -- should be below 64KB (consider reduction of page size & use of GZIP to compress files)

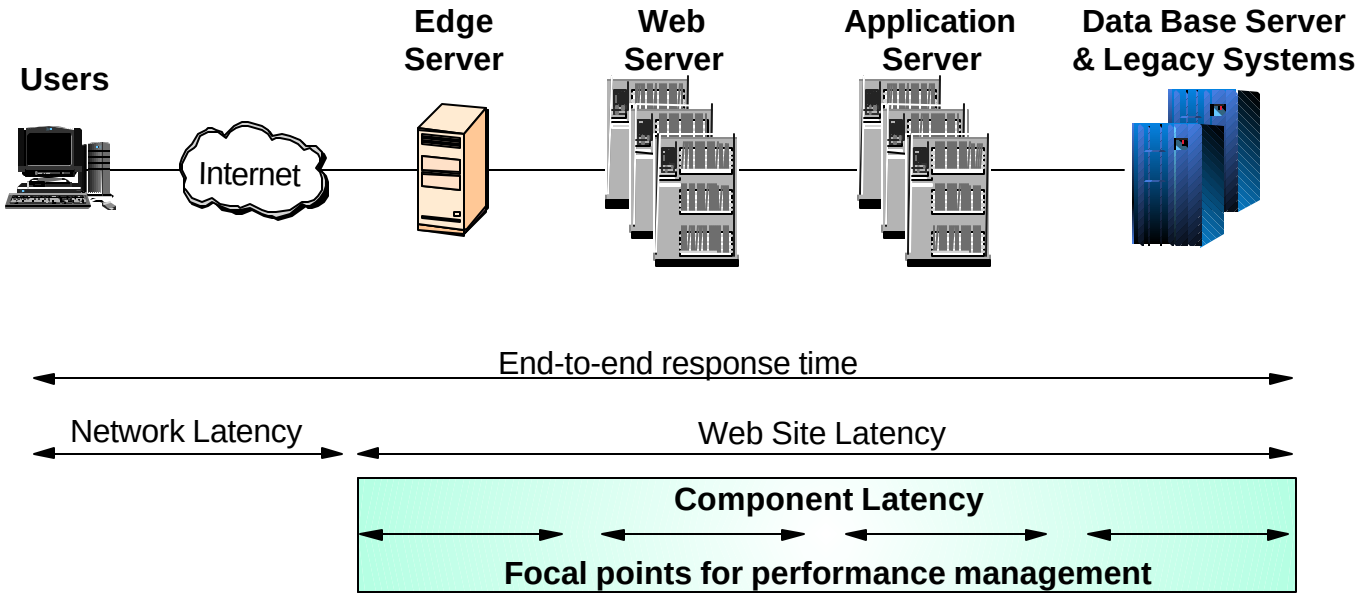
- 50 items requested -- should generally be <= 20
- Considerable overhead to request/respond/deliver
- Considerable overhead to establish/close 37 connections
- Consider consolidation of items where possible

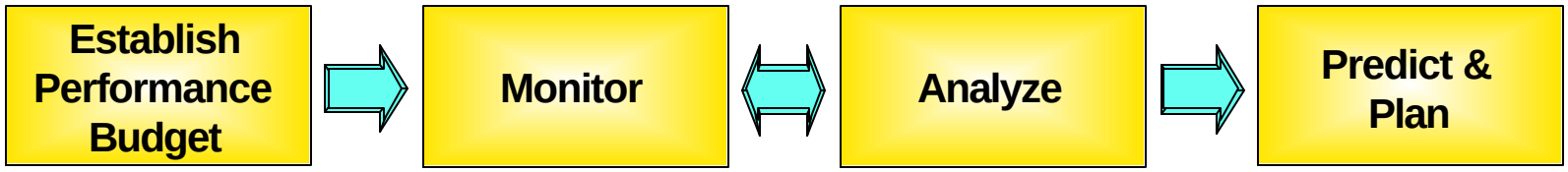
- High server response times (0.5-2.5 seconds)
- Objective is 0.5 seconds

Manage End-to-End Performance Lessons Learned

- Use proactive end-to-end measurements to identify constraints and bottlenecks
- Consider performance management methodology from the beginning
  - e-Business performance requirements
  - New Web site architecures
- Enforce processes that link performance objectives to planning, design and operations

e-Business Web Application System
End-To-End Performance Management Methodology





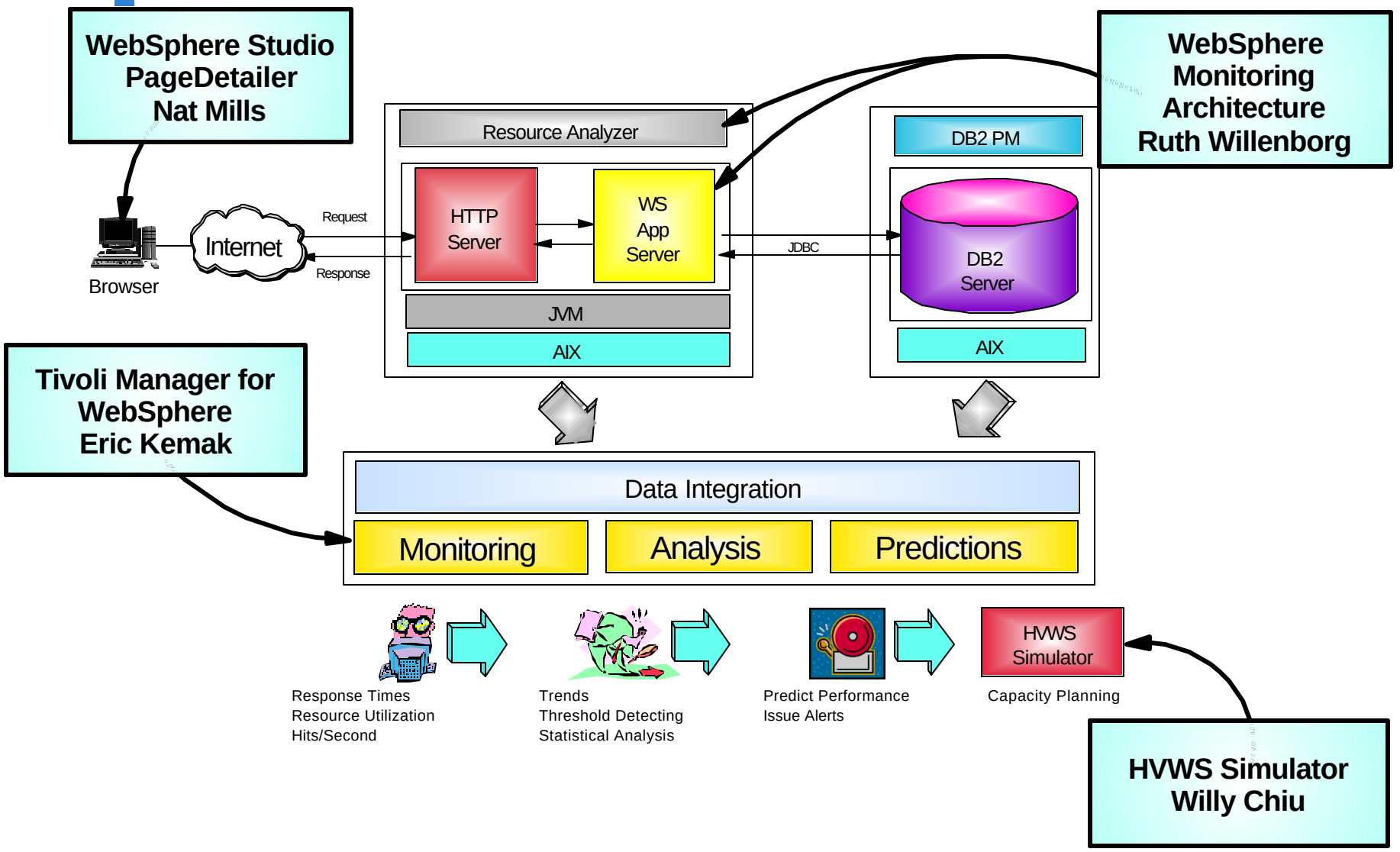
Methodology Overview

HVWS White Papers

www.ibm.com/developerworks/library/hvws/

- Design for Scalability
  - Describes component selection and management techniques you can use to make your Web site ready to adapt to increasing traffic
- Planning for Growth
  - Introduces a methodology for capacity planning, identifying workload patterns and configuring site infrastructure
- Design Pages for Performance
  - Discusses what drives Web page download times
  - Introduces page design practices that can reduce download time and improve resource utilization
- Manage Web Site Performance
  - Discusses end to end performance management
- Web Site Personalization
  - Introduces current and future techniques for personalizing your Web site

Tivoli & WebSphere Demonstrations



SESSION **"Best Practices for High Volume Web Sites"**

NUMBER *Presentation Number PE420511, E420521*

SPEAKER **Dr. Willy Chiu**

TITLE **VP, High Volume Web Sites**

AFFILIATION **IBM**

wchiu@us.ibm.com